

Achieving Non-Discrimination in Prediction

Lu Zhang, Yongkai Wu, and Xintao Wu

University of Arkansas

{lz006,yw009,xintaowu}@uark.edu

Abstract

Discrimination-aware classification is receiving an increasing attention in the data mining and machine learning fields. The data preprocessing methods for constructing a discrimination-free classifier remove discrimination from the training data, and learn the classifier from the cleaned data. However, there lacks of a theoretical guarantee for the performance of these methods. In this paper, we fill this theoretical gap by mathematically bounding the probability that the discrimination in predictions is within a given interval in terms of the given training data and classifier. In our analysis, we adopt the causal model for modeling the mechanisms in data generation, and formally defining discrimination in the population, in a dataset, and in the prediction. The theoretical results show that the fundamental assumption made by the data preprocessing methods is not correct. Finally, we develop a framework for constructing a discrimination-free classifier with a theoretical guarantee.

1 Introduction

Discrimination-aware classification is receiving an increasing attention in the data mining and machine learning fields. Classifiers are learned from the training data to reduce the gap between the predicted labels and the true labels indicated by the data. If the training data is discriminatory, the predictions made by the classifiers may also contain discrimination. Therefore, it is of great interest to constructing a discrimination-free classifier, i.e., there will be no discrimination in prediction for the unlabeled new data, even if the training data is discriminatory.

A large family of proposed methods for constructing discrimination-free classifiers, referred to as the data preprocessing methods, are based on removing discrimination from the training dataset, and then learning the classifier on the cleaned dataset. The fundamental assumption for these methods is that, since the classifier is learned from a discrimination-free dataset, it is likely that the future predictions will also be “more” discrimination-free [Kamiran and Calders, 2009b]. Although this assumption is plausible, however, there is no theoretical guarantee to show

“how much likely” and “how” discrimination-free the predictions would be given a training data and a classifier. There is also no theoretical guarantee to show “how” discrimination-free the predictions would be if the classifier is learned from a modified dataset but the unlabeled new data for prediction are drawing from the original population. The lack of the theoretical guarantees places great uncertainty on the performance of all the data preprocessing methods.

In this paper, we fill the above theoretical gap by mathematically bounding the probability that the discrimination in predictions is within a given interval in terms of the given training data and classifier. We obtain two important theoretical results: 1) even when discrimination in the training data is completely removed, the prediction can still contain non-negligible amount of discrimination, caused by the bias in the classifier; and 2) for removing discrimination, different from the claims of many previous work (e.g., [Feldman *et al.*, 2015]), not all methods can ensure non-discrimination in predictions even though they can achieve non-discrimination on the modified training data. Based on the results, we develop a two-phase framework for constructing a discrimination-free classifier with a theoretical guarantee.

In our analysis, we adopt the causal model for modeling the mechanisms in data generation, and formally defining discrimination in the population, in a dataset, and in the prediction. A causal model [Pearl, 2009] is a structural equation-based mathematical object that describes the causal mechanisms of a system. It is widely assumed in the machine learning field that there exists a fixed but unknown data population where both the training data and the unlabeled new data are drawn from. We further assume that there exists a fixed but unknown causal model that represents the data generation mechanisms of the population. By using the causal model, we formally define discrimination as the causal effects of the protected attribute on the label, and derive the formula for quantitatively measuring the discriminatory effect from the conditional probabilities in the data population. We then derive the discriminatory effect in a dataset, as well as the discriminatory effect in the prediction. Finally, we link the discrimination in the prediction with the discrimination in the training data by a probabilistic condition, which provides a guideline to achieve non-discrimination in the prediction by employing the existing preprocessing methods.

Table 1: Table of notations.

Notation	Definition
C	protected attribute
$\mathbf{R} = \{R_1, \dots, R_m\}$	non-protected attributes
L	label
$h : C \times \mathbf{R} \rightarrow L$	classifier
\mathcal{M}	causal model or mechanisms
\mathcal{M}_h	causal model of prediction
$\mathcal{D} = \{(c^{(j)}, \mathbf{r}^{(j)}, l^{(j)})\}$	training data
$\mathcal{D}_h = \{(c^{(j)}, \mathbf{r}^{(j)}, h(c^{(j)}, \mathbf{r}^{(j)}))\}$	prediction over \mathcal{D}

2 Preliminary Concepts

2.1 Notations and Representations

We consider an attribute space which consists of some protected attributes, the label, and the non-protected attributes. Throughout the paper, we use an uppercase alphabet, e.g., X to represent an attribute; a bold uppercase alphabet, e.g., \mathbf{X} , to represent a subset of attributes. We use a lowercase alphabet, e.g., x , to represent a realization or instantiation of attribute X ; a bold lowercase alphabet, e.g., \mathbf{x} , to represent a realization or instantiation of \mathbf{X} . For ease of representation, we assume that there is only one protected attribute, denoted by C , which is a binary attribute associated with the domain values of the non-protected group c^+ and the protected group c^- . We denote the label by L , which is a binary attribute associated with the domain values of the positive label l^+ and negative label l^- . According to the convention in machine learning, we also define that $l^+ = 1$ and $l^- = 0$. The set of all the non-protected attributes is denoted by $\mathbf{R} = \{R_1, \dots, R_m\}$. Please refer to the notation table shown as Table 1.

2.2 Causal Model

A causal model is a mathematical object that describes the causal mechanisms of a system as a set of structural equations. It is formally defined as follows.

Definition 1 (Causal Model). *A causal model is a triple $\mathcal{M} = (\mathbf{U}, \mathbf{V}, \mathbf{F})$ where*

1. \mathbf{U} is a set of arbitrarily distributed random variables (called *exogenous*) that are determined by factors outside the model.
2. A joint probability distribution $P(\mathbf{u})$ is defined over the variables in \mathbf{U} .
3. \mathbf{V} is a set $\{X_1, \dots, X_i, \dots\}$ of variables (called *endogenous*) that are determined by variables in the model, namely, variables in $\mathbf{U} \cup \mathbf{V}$.
4. \mathbf{F} is a set of deterministic functions $\{f_1, \dots, f_i, \dots\}$ where each f_i is a mapping from $\mathbf{U} \times (\mathbf{V} \setminus X_i)$ to X_i . Symbolically, the set of equations \mathbf{F} can be represented by writing

$$x_i = f_i(pa_i, u_i)$$

where pa_i is any realization of the unique minimal set of variables PA_i in $\mathbf{V} \setminus X_i$ (referred to as the *parents* of X_i) that renders f_i nontrivial. Similarly, $U_i \subset \mathbf{U}$ stands for the unique minimal set of variables in \mathbf{U} that renders f_i nontrivial.

Each model \mathcal{M} is associated with a direct graph $\mathcal{G}(\mathcal{M})$, where each node in the graph corresponds to a variable X_i in \mathbf{V} , and direct edges point from each member of PA_i toward X_i . Such graph is called the causal graph associated with \mathcal{M} .

The causal model generalizes naturally to probabilistic systems, as shown in the following relationship. For each variable $Y \in \mathbf{V}$, denote the value of Y given an instantiation $\mathbf{U} = \mathbf{u}$ by $Y(\mathbf{u})$. Then it follows that

$$P(y) \triangleq P(Y = y) = \sum_{\{\mathbf{u}: Y(\mathbf{u})=y\}} P(\mathbf{u}).$$

The causal effect in the causal model is defined over an intervention that fixes the value of an endogenous variable(s) X to a constant(s) x . The intervention is achieved by deleting the variable X and its associated function from the model, replacing them with the constant x , while keeping the rest of the model unchanged, which is mathematically formalized as $do(X = x)$ or simply $do(x)$. For any variables $X, Y \in \mathbf{V}$, denote the value of Y under $do(x)$ given an instantiation $\mathbf{U} = \mathbf{u}$ by $Y_x(\mathbf{u})$. Then, the causal effect of X on Y is defined as

$$P(y|do(x)) \triangleq P(Y = y|do(X = x)) = \sum_{\{\mathbf{u}: Y_x(\mathbf{u})=y\}} P(\mathbf{u}). \quad (1)$$

One strength of the causal model is that, the causal effect $P(y|do(x))$ can be computed from the traditional probabilities, under some common assumptions. One class of such causal models is called Markovian. A causal model is said to be Markovian if: 1) its associated causal graph is acyclic; and 2) all variables in \mathbf{U} are mutually independent. The above two requirements are equivalent to the parental Markov condition [Koller and Friedman, 2009], which is the fundamental assumption in the probabilistic graphical models. The following theorem shows how $P(y|do(x))$ is computed in a Markovian model.

Theorem 1 (Truncated Factorization). *For any Markovian model, the causal effect $P(y|do(x))$ over two endogenous variables X and Y is given by the truncated factorization*

$$P(y|do(x)) = \sum_{\mathbf{V} \setminus \{X, Y\}} \prod_{Y=y} \prod_{X_i \neq X} P(x_i | pa_i) \delta_{X=x},$$

where the summation is a marginalization that traverses all value combinations of $\mathbf{V} \setminus \{X, Y\}$, and $\delta_{X=x}$ means replacing X with x in each term.

3 Discrimination in Population and Dataset

We formally define discrimination using the causal model. Assume that there exists a fixed but unknown population over the space $C \times \mathbf{R} \times L$, and there exists a fixed but unknown causal model \mathcal{M} representing the mechanisms that determine the values of all the attributes in the population. Without ambiguity, we also use \mathcal{M} to denote the population, and the terms mechanisms and population are used interchangeably. We assume that \mathcal{M} is Markovian. We further make two reasonable assumptions under our context: 1) the protected attribute C has no parent in \mathbf{V} ; and 2) the label L has no child in \mathbf{V} . Then, the causal model can be written as follows.

$$\begin{aligned} \text{Model } \mathcal{M} \quad & c = f_C(u_C) \\ & r_i = f_i(pa_i, u_i) \quad i = 1, \dots, m \\ & l = f_L(pa_L, u_L) \end{aligned}$$

We first define discrimination on \mathcal{M} , which can be considered as the true discrimination existed in the data generation mechanisms. The central question of discrimination asks, whether the label of an individual would be different had the individual been of a different protected group (e.g., sex, race, age, religion, etc.). Note that when an instantiation $\mathbf{U} = \mathbf{u}$ is given, the causal model is completely specified at the individual level. For each individual specified by \mathbf{u} , we consider the difference in the labels he/she would receive if we intervene his/her value of protected attribute C . The label when C is fixed to c^+ is given by $L_{c^+}(\mathbf{u})$, and the label when C is fixed to c^- is given by $L_{c^-}(\mathbf{u})$. Thus, the difference in the labels of the individual is given by

$$L_{c^+}(\mathbf{u}) - L_{c^-}(\mathbf{u}).$$

The expected change across all individuals is hence given by

$$\mathbb{E}[L_{c^+}(\mathbf{u}) - L_{c^-}(\mathbf{u})].$$

We define this expected change as the (average) discriminatory effect in \mathcal{M} , denoted by $\text{DE}(c^+, c^-)_{\mathcal{M}}$.

Definition 2. The discriminatory effect in a causal model \mathcal{M} is given by

$$\text{DE}(c^+, c^-)_{\mathcal{M}} = \mathbb{E}[L_{c^+}(\mathbf{u}) - L_{c^-}(\mathbf{u})].$$

Note that the effect of the reverse discrimination can be similarly given by $\text{DE}(c^-, c^+)_{\mathcal{M}}$. Therefore, given a user-defined threshold τ ($\tau \geq 0$), the definition of discrimination in the mechanisms or population is given as follows.

Definition 3. Given a causal model \mathcal{M} and a threshold τ ($\tau \geq 0$), discrimination exists if either $\text{DE}(c^+, c^-)_{\mathcal{M}} > \tau$ or $\text{DE}(c^-, c^+)_{\mathcal{M}} > \tau$ holds.

The following theorem shows how $\text{DE}(c^+, c^-)_{\mathcal{M}}$ is computed from the population, given that \mathcal{M} is Markovian.

Theorem 2. Given a causal model \mathcal{M} , the discriminatory effect in \mathcal{M} is computed by

$$\text{DE}(c^+, c^-)_{\mathcal{M}} = P(I^+|c^+) - P(I^+|c^-).$$

Proof. The expectation is represented by

$$\begin{aligned} \mathbb{E}[L_{c^+}(\mathbf{u})] &= \sum_{\mathbf{u}} L_{c^+}(\mathbf{u})P(\mathbf{u}) \\ &= \sum_{\{\mathbf{u}: L_{c^+}(\mathbf{u})=I^+\}} I^+ P(\mathbf{u}) + \sum_{\{\mathbf{u}: L_{c^+}(\mathbf{u})=I^-\}} I^- P(\mathbf{u}) = \sum_{\{\mathbf{u}: L_{c^+}(\mathbf{u})=I^+\}} P(\mathbf{u}). \end{aligned} \quad (2)$$

According to Equation (1), the above expression equals $P(I^+|do(c^+))$. According to Theorem 1 and the assumption that C has no parent, it is straightforward to derive that $P(I^+|do(c^+)) = P(I^+|c^+)$. Thus, we have $\mathbb{E}[L_{c^+}(\mathbf{u})] = P(I^+|c^+)$. Similarly we can prove $\mathbb{E}[L_{c^-}(\mathbf{u})] = P(I^+|c^-)$. Hence, the theorem is proven. \square

Theorem 2 shows that $\text{DE}(c^+, c^-)_{\mathcal{M}} = -\text{DE}(c^-, c^+)_{\mathcal{M}}$. In the following, we simplify $\text{DE}(c^+, c^-)_{\mathcal{M}}$ to $\text{DE}_{\mathcal{M}}$, and the conditions in Definition 3 is simplified to $|\text{DE}_{\mathcal{M}}| > \tau$.

Interestingly, our obtained discrimination measurement is the same as the classic discrimination metric *risk difference*, which is widely used as the non-discrimination constraint in

discrimination-aware learning [Romei and Ruggieri, 2014]. Thus, our analysis can help understand the assumptions and scenarios in which the risk difference applies.

In practice, \mathcal{M} is unknown and we can only observe a dataset $\mathcal{D} = \{(c^{(j)}, \mathbf{r}^{(j)}, I^{(j)}); j = 1, \dots, n\}$ sampled from the population. We define the discrimination on \mathcal{D} , denoted by $\text{DE}_{\mathcal{D}}$, as the maximum likelihood estimation of $\text{DE}_{\mathcal{M}}$.

Proposition 1. Given a dataset \mathcal{D} , the discriminatory effect in \mathcal{D} is given by

$$\text{DE}(c^+, c^-)_{\mathcal{D}} = \hat{P}(I^+|c^+) - \hat{P}(I^+|c^-),$$

where $\hat{P}(\cdot)$ s are the conditional frequencies in \mathcal{D} .

To estimate $\text{DE}_{\mathcal{M}}$ using $\text{DE}_{\mathcal{D}}$, we bound the distance between $\text{DE}_{\mathcal{M}}$ and $\text{DE}_{\mathcal{D}}$ in term of the sample size of \mathcal{D} .

Proposition 2. For any dataset \mathcal{D} with size of n generated by a causal model \mathcal{M} , the probability of that the distance between $\text{DE}_{\mathcal{M}}$ and $\text{DE}_{\mathcal{D}}$ is no larger than t is bounded by

$$P(|\text{DE}_{\mathcal{M}} - \text{DE}_{\mathcal{D}}| \leq t) \geq 1 - \sqrt{\frac{8\pi}{n}} e^{-\frac{2n^+n^-}{n}t^2},$$

where n^+ and n^- ($n^+ + n^- = n$) are the numbers of individuals with c^- and c^+ in \mathcal{D} .

Proof. By definition of $\text{DE}_{\mathcal{M}}$ and $\text{DE}_{\mathcal{D}}$ we have

$$\text{DE}_{\mathcal{M}} - \text{DE}_{\mathcal{D}} = (P(I^+|c^+) - \hat{P}(I^+|c^+)) + (\hat{P}(I^+|c^-) - P(I^+|c^-)).$$

Denoting by $I^{(+j)}$ the label of the j th individual in \mathcal{D} with $C = c^+$, we can write $\hat{P}(I^+|c^+)$ as

$$\hat{P}(I^+|c^+) = \frac{1}{n^+} (\mathbb{1}_{[I^{(+1)}=I^+]} + \dots + \mathbb{1}_{[I^{(n^+)}=I^+]}),$$

where indicators $\mathbb{1}_{[I^{(+j)}=I^+]}$ ($j = 1 \dots n^+$) can be considered as independent random variables bounded by the interval $[0, 1]$. Note that $\mathbb{E}[\hat{P}(I^+|c^+)] = P(I^+|c^+)$. According to the Hoeffding's inequality [Murphy, 2012], we have

$$P(|P(I^+|c^+) - \hat{P}(I^+|c^+)| \geq t) \leq 2e^{-2n^+t^2}.$$

Similarly, we have

$$P(|P(I^+|c^-) - \hat{P}(I^+|c^-)| \geq t) \leq 2e^{-2n^-t^2}.$$

Therefore, we have

$$\begin{aligned} &P(|\text{DE}_{\mathcal{M}} - \text{DE}_{\mathcal{D}}| \geq t) \\ &\leq P(|P(I^+|c^+) - \hat{P}(I^+|c^+)| + |P(I^+|c^-) - \hat{P}(I^+|c^-)| \geq t) \\ &= \int_0^1 P(|P(I^+|c^+) - \hat{P}(I^+|c^+)| \geq x \wedge |P(I^+|c^-) - \hat{P}(I^+|c^-)| \geq t-x) dx \\ &= \int_0^1 P(|P(I^+|c^+) - \hat{P}(I^+|c^+)| \geq x) P(|P(I^+|c^-) - \hat{P}(I^+|c^-)| \geq t-x) dx \\ &\leq \int_0^1 4e^{-2n^+x^2} e^{-2n^-(t-x)^2} dx \leq \sqrt{\frac{8\pi}{n}} e^{-\frac{2n^+n^-}{n}t^2}. \end{aligned}$$

The fourth line of the above expression is due to that each individual is independently drawn from the population. \square

Table 2: The confusion matrix.

	Predicted positive	Predicted negative
True positive	tp	fn
True negative	fp	tn

4 Discrimination in Prediction

So far we have not introduced the classifier. In this section, we estimate the discrimination in the predictions made by the classifier. In the learning theorem, a classifier h is function mapping from $C \times \mathbf{R}$ to L , i.e., $h : C \times \mathbf{R} \rightarrow L$. The space of functions \mathcal{H} is the set of candidate functions. A learning algorithm analyzes the training dataset \mathcal{D} to find a function from \mathcal{H} that minimizes the difference between the predicted labels $h(c^{(j)}, \mathbf{r}^{(j)})$ and the true labels $l^{(j)}$ ($j = 1, \dots, m$). The training performance of a classifier is characterized by the confusion matrix shown in Table 2. Specifically, fp/n is known as the false positive rate, which we denote by ε_1 , and fn/n is known as the false negative rate, which we denote by ε_2 .

Once training completes, the classifier is deployed to infer predictions on the unlabeled new data, i.e., the classifier computes the predicted label for any unlabeled individual. We can assume that the unlabeled data is drawn from the same population as the training data, i.e., from \mathcal{M} , with the labels unknown. Therefore, in the predictions, the values of all the attributes other than the label are still determined by the mechanisms in \mathcal{M} , while the classifier now acts as a new mechanism for determining the value of the label. We consider the mechanisms from \mathcal{M} over the sets of variables \mathbf{U} and \mathbf{V} with function $f_L(\cdot)$ replaced with classifier $h(\cdot)$ as a new causal model, denoted by \mathcal{M}_h . It is written as

$$\begin{aligned} \text{Model } \mathcal{M}_h \quad & c = f_C(u_C) \\ & r_i = f_i(pa_i, u_i) \quad i = 1, \dots, m \\ & l = h(c, \mathbf{r}) \end{aligned}$$

In this way, discrimination in prediction is given by the discrimination in \mathcal{M}_h , denoted by $\text{DE}_{\mathcal{M}_h}$. In addition, it is clear that \mathcal{M}_h is also Markovian. Hence, similar to Theorem 2, the computation of $\text{DE}_{\mathcal{M}_h}$ is given as follows.

Proposition 3. *Given a causal model \mathcal{M} and a classifier h , the discriminatory effect in \mathcal{M}_h is given by*

$$\text{DE}_{\mathcal{M}_h} = P(h^+|c^+) - P(h^+|c^-),$$

where $P(h^+|c^+)$ (and similarly for $P(h^+|c^-)$) is the probability of the classifier to predict positive labels for the data with $C = c^+$, given by

$$P(h^+|c^+) = \sum_{\mathbf{R}} \mathbb{1}_{[h(c^+, \mathbf{r})=l^+]} P(\mathbf{r}|c^+).$$

Proof. It is directly extended from Theorem 2 that

$$\text{DE}_{\mathcal{M}_h} = P(h^+|c^+) - P(h^+|c^-).$$

According to Equation (2), we have

$$\begin{aligned} P(h^+|c^+) &= \mathbb{E}[L_{c^+}(\mathbf{u})] = \sum_{\{\mathbf{u}: L_{c^+}(\mathbf{u})=l^+\}} P(\mathbf{u}) = \sum_{\mathbf{R}} \sum_{\{\mathbf{u}: \mathbf{R}_{c^+}(\mathbf{u})=\mathbf{r}, \\ &\quad h(c^+, \mathbf{r})=l^+\}} P(\mathbf{u}) \\ &= \sum_{\mathbf{R}} \mathbb{1}_{[h(c^+, \mathbf{r})=l^+]} \sum_{\{\mathbf{u}: \mathbf{R}_{c^+}(\mathbf{u})=\mathbf{r}\}} P(\mathbf{u}) = \sum_{\mathbf{R}} \mathbb{1}_{[h(c^+, \mathbf{r})=l^+]} P(\mathbf{r}|do(c^+)). \end{aligned}$$

Similarly, we can derive that $P(\mathbf{r}|do(c^+)) = P(\mathbf{r}|c^+)$. Hence, the theorem is proven. \square

Note that \mathcal{M}_h is also unknown. To estimate $\text{DE}_{\mathcal{M}_h}$, we apply the classifier on the training data \mathcal{D} , and obtain a new dataset \mathcal{D}_h by replacing the original labels with the predicted labels, i.e., $\mathcal{D}_h = \{(c^{(j)}, \mathbf{r}^{(j)}, h(c^{(j)}, \mathbf{r}^{(j)}))\}; j = 1, \dots, n\}$. Thus, \mathcal{D}_h can be considered as a sample drawn from \mathcal{M}_h . Then, we similarly define the discrimination on \mathcal{D}_h as the maximum likelihood estimation of $\text{DE}_{\mathcal{M}_h}$ as follows.

Proposition 4. *Given a dataset \mathcal{D} and a classifier h , the discriminatory effect in \mathcal{D}_h is given by*

$$\text{DE}_{\mathcal{D}_h} = \hat{P}(h^+|c^+) - \hat{P}(h^+|c^-),$$

where

$$\hat{P}(h^+|c^+) = \sum_{\mathbf{R}} \mathbb{1}_{[h(c^+, \mathbf{r})=l^+]} \hat{P}(\mathbf{r}|c^+) = \frac{1}{n^+} \sum_{\{j: c^{(j)}=c^+\}} \mathbb{1}_{[h(c^{(j)}, \mathbf{r}^{(j)})=l^+]}. \quad \text{Here } n^+ \text{ and } n^- \text{ (} n^+ + n^- = n \text{) are the numbers of individuals with } c^- \text{ and } c^+ \text{ in } \mathcal{D}.$$

Similar to Proposition 2, we bound the distance between $\text{DE}_{\mathcal{M}_h}$ and $\text{DE}_{\mathcal{D}_h}$ in term of the sample size of \mathcal{D} . The following proposition is directly extended from Proposition 2.

Proposition 5. *For any dataset \mathcal{D} with size of n generated by a causal model \mathcal{M} , and any classifier $h : C \times \mathbf{R} \rightarrow L$, the probability of that the distance between $\text{DE}_{\mathcal{M}_h}$ and $\text{DE}_{\mathcal{D}_h}$ is no larger than t is bounded by*

$$P\left(|\text{DE}_{\mathcal{M}_h} - \text{DE}_{\mathcal{D}_h}| \leq t\right) \geq 1 - \sqrt{\frac{8\pi}{n}} e^{-\frac{2n^+n^-}{n} t^2}.$$

Next, we give the relation between discrimination in classifier training and discrimination in training data, i.e., $\text{DE}_{\mathcal{D}_h}$ and $\text{DE}_{\mathcal{D}}$, in term of the training performance of the classifier. The performance measure we used is what we refer to as the *error bias*.

Definition 4 (Error Bias). *For any classifier that is learned from a training data \mathcal{D} , the error bias is given by*

$$\varepsilon_{h, \mathcal{D}} = \varepsilon_1^+ - \varepsilon_2^+ - (\varepsilon_1^- - \varepsilon_2^-),$$

where $\varepsilon_1^+, \varepsilon_1^-$ are the false positive rates on data with $C = c^+$ and $C = c^-$ respectively, and $\varepsilon_2^+, \varepsilon_2^-$ are the false negative rates on data with $C = c^+$ and $C = c^-$ respectively.

Proposition 6. *For any classifier h that is learned from \mathcal{D} , we have*

$$\text{DE}_{\mathcal{D}_h} - \text{DE}_{\mathcal{D}} = \varepsilon_{h, \mathcal{D}}.$$

Proof. By definition, the false positive rate on data with $C = c^+$ is given by

$$\varepsilon_1^+ = \frac{1}{n^+} \sum_{\{j: c^{(j)}=c^+, l^{(j)}=l^+\}} \mathbb{1}_{[h(c^{(j)}, \mathbf{r}^{(j)})=l^+]},$$

which can be rewritten as

$$\varepsilon_1^+ = \sum_{\mathbf{R}} \hat{P}(\mathbf{r}|c^+) \cdot \mathbb{1}_{[h(c^{(j)}, \mathbf{r}^{(j)})=l^+]} \cdot (1 - \hat{P}(l^+|c^+, \mathbf{r})).$$

Similarly, the false negative rate on data with $C = c^+$ can be given by

$$\varepsilon_2^+ = \sum_{\mathbf{r}} \hat{P}(\mathbf{r}|c^+) \cdot \mathbb{1}_{[h(c^+), \mathbf{r}^{(j)}] = l^-] \cdot \hat{P}(l^+|c^+, \mathbf{r}).$$

Subtracting ε_2^+ from ε_1^+ , we obtain

$$\varepsilon_1^+ - \varepsilon_2^+ = \sum_{\mathbf{r}} \hat{P}(\mathbf{r}|c^+) \left(\mathbb{1}_{[h(c^+), \mathbf{r}] = l^+]} (1 - \hat{P}(l^+|c^+, \mathbf{r})) - \mathbb{1}_{[h(c^+), \mathbf{r}] = l^-]} \hat{P}(l^+|c^+, \mathbf{r}) \right),$$

which is equivalent to

$$\varepsilon_1^+ - \varepsilon_2^+ = \sum_{\mathbf{r}} \hat{P}(\mathbf{r}|c^+) \cdot \left(\mathbb{1}_{[h(c^+), \mathbf{r}] = l^+]} - \hat{P}(l^+|c^+, \mathbf{r}) \right).$$

Similarly for data with $C = c^-$, we have

$$\varepsilon_1^- - \varepsilon_2^- = \sum_{\mathbf{r}} \hat{P}(\mathbf{r}|c^-) \cdot \left(\mathbb{1}_{[h(c^-), \mathbf{r}] = l^+]} - \hat{P}(l^+|c^-, \mathbf{r}) \right).$$

It follows that

$$\begin{aligned} \text{DE}_{\mathcal{D}_h} - \text{DE}_{\mathcal{D}} &= \sum_{\mathbf{r}} \hat{P}(\mathbf{r}|c^+) \left(\mathbb{1}_{[h(c^+), \mathbf{r}] = l^+]} - \hat{P}(l^+|c^+, \mathbf{r}) \right) \\ &\quad - \sum_{\mathbf{r}} \hat{P}(\mathbf{r}|c^-) \left(\mathbb{1}_{[h(c^-), \mathbf{r}] = l^+]} - \hat{P}(l^+|c^-, \mathbf{r}) \right) = \varepsilon_1^+ - \varepsilon_2^+ - (\varepsilon_1^- - \varepsilon_2^-). \end{aligned}$$

Let $\varepsilon_{h, \mathcal{D}} = \varepsilon_1^+ - \varepsilon_2^+ - (\varepsilon_1^- - \varepsilon_2^-)$ completes the proof. \square

Combining Propositions 5 and 6, we bound $\text{DE}_{\mathcal{M}_h}$ in terms of $\text{DE}_{\mathcal{D}}$, as well as $\varepsilon_{h, \mathcal{D}}$ and n .

Theorem 3. *For any dataset \mathcal{D} with size of n generated by \mathcal{M} , and any classifier h trained on \mathcal{D} , given a user-defined parameter τ ($\tau \geq 0$), if $|\text{DE}_{\mathcal{D}} + \varepsilon_{h, \mathcal{D}}| \leq \tau$, then we have*

$$P \left(|\text{DE}_{\mathcal{M}_h}| \leq \tau + t \right) \geq 1 - \sqrt{\frac{8\pi}{n}} e^{-\frac{2n^+n^-}{n} t^2}.$$

Theorem 3 shows that, the discrimination in the training data and the error bias in the classifier are both factors that will determine the discrimination in predictions. Given a discrimination-free dataset \mathcal{D} , i.e., $|\text{DE}_{\mathcal{D}}| \leq \tau$, we cannot guarantee that any classifier learned from it would not produce discriminatory predictions. To ensure discrimination-free predictions with high probability, we must ensure that the sum of $\text{DE}_{\mathcal{D}}$ and $\varepsilon_{h, \mathcal{D}}$ are within the given interval.

5 Achieve Non-Discrimination in Prediction

This section solves the key question in data preprocessing methods: if the training data contains discrimination, can we achieve non-discrimination in prediction through removing discrimination from the training data? In [Feldman *et al.*, 2015], the authors claim non-discrimination for the prediction. However, their claim is based on the modified training data, not on the future predictions.

The following theorem shows that the answer is guaranteed to be yes when only the labels of \mathcal{D} are modified during the modifying process.

Theorem 4. *For any dataset \mathcal{D} with size of n generated by \mathcal{M} , let \mathcal{D}^* be a dataset obtained from \mathcal{D} by only modifying its labels. Given a user-defined parameter τ ($\tau \geq 0$), for any new classifier h^* trained on \mathcal{D}^* , if $|\text{DE}_{\mathcal{D}^*} + \varepsilon_{h^*, \mathcal{D}^*}| \leq \tau$, then we have*

$$P \left(|\text{DE}_{\mathcal{M}_{h^*}}| \leq \tau + t \right) \geq 1 - \sqrt{\frac{8\pi}{n}} e^{-\frac{2n^+n^-}{n} t^2}.$$

Proof. Let \mathcal{M}^* be the causal model that generates \mathcal{D}^* . According to Theorem 3, it follows that

$$P \left(|\text{DE}_{\mathcal{M}_{h^*}^*}| \leq \tau + t \right) \geq 1 - \sqrt{\frac{8\pi}{n}} e^{-\frac{2n^+n^-}{n} t^2}.$$

The key to the proof is to show that $\mathcal{M}_{h^*}^* \triangleq \mathcal{M}_{h^*}$. Causal model \mathcal{M}_{h^*} can be written as

$$\begin{aligned} \text{Model } \mathcal{M}_{h^*} \quad & c = f_C(u_C) \\ & r_i = f_i(pa_i, u_i) \quad i = 1, \dots, m \\ & l = h^*(c, \mathbf{r}) \end{aligned}$$

Since \mathcal{D}^* is obtained from \mathcal{D} by only modifying L , without loss of generality, we can assume that the causal model \mathcal{M}^* that generates \mathcal{D}^* is different from \mathcal{M} only in the function that determines L , while the generation of all other variables remain unchanged. Thus, we can write \mathcal{M}^* as follows

$$\begin{aligned} \text{Model } \mathcal{M}^* \quad & c = f_C(u_C) \\ & r_i = f_i(pa_i, u_i) \quad i = 1, \dots, m \\ & l = f_L^*(pa_L^*, u_L^*) \end{aligned}$$

Then, $\mathcal{M}_{h^*}^*$ is given by

$$\begin{aligned} \text{Model } \mathcal{M}_{h^*}^* \quad & c = f_C(u_C) \\ & r_i = f_i(pa_i, u_i) \quad i = 1, \dots, m \\ & l = h^*(c, \mathbf{r}) \end{aligned}$$

Thus, we have $\mathcal{M}_{h^*}^* \triangleq \mathcal{M}_{h^*}$, which completes the proof. \square

On the other hand, if any attribute other than L is modified when removing discrimination from the training data, we cannot obtain the bounded guarantee for non-discrimination in the prediction. We will show using an empirical example in the next section that, even when the classifier is built on a discrimination-free dataset, and the error bias in the classifier is also removed, there still exists discrimination in the prediction. The intuition behind the results is simple. If we only modify the labels, the modified training data will have no inconsistency with the new data since the new data is unlabeled. However, if we modify the attributes other than L , we can obtain the discrimination-free prediction based on the modified data. Nevertheless, the new data drawn from the original population is inconsistent with the modified data, hence the discrimination-free prediction does not apply to the new data.

5.1 Two-Phase Modifying

Theorem 4 provides a guideline to achieve non-discrimination in the prediction, which shows that we may need to modify the training data to reduce its discrimination, and also modify the classifier to reduce the error bias, in order to achieve $|\text{DE}_{\mathcal{D}^*} + \varepsilon_{h^*, \mathcal{D}^*}| \leq \tau$. It should be noted that, when the training data is changed, the error bias of the

Algorithm 1: Two-phase framework.

- 1 If $|\text{DE}_{\mathcal{D}^*} + \varepsilon_{h^*, \mathcal{D}^*}| \leq \tau$, we are done. Otherwise, modify the labels in the training dataset \mathcal{D} to obtain a modified dataset \mathcal{D}^* such that $|\text{DE}_{\mathcal{D}^*}| \leq \tau$.
 - 2 Train a classifier h^* on \mathcal{D}^* . If $|\text{DE}_{\mathcal{D}^*} + \varepsilon_{h^*, \mathcal{D}^*}| \leq \tau$, we are done. Otherwise, modify classifier h^* to meet the above requirement.
-

classifier built on it will also change. Therefore, we propose a two-phase framework for modifying the training data as well as the classifier, shown in Algorithm 1

Suppose that $|\text{DE}_{\mathcal{D}^*} + \varepsilon_{h^*, \mathcal{D}^*}| > \tau$. In the first phase, if $|\text{DE}_{\mathcal{D}}| > \tau$, then we modify \mathcal{D} to reduce the discrimination it contains. A number of methods have been proposed for removing discrimination from the training data, which can be generally divided into two categories: 1) those that only modify the labels of the data; and 2) those that modify the attributes other than the label, i.e., C and \mathbf{R} . As stated above, only the methods from the first category that can achieve $|\text{DE}_{\mathcal{D}}| \leq \tau$ for the modified dataset \mathcal{D}^* can ensure non-discrimination in the prediction using our framework, while the methods from the second class cannot.

In the second phase, after learning a classifier h^* from \mathcal{D}^* , if we still have $|\text{DE}_{\mathcal{D}^*} + \varepsilon_{h^*, \mathcal{D}^*}| > \tau$, then we make some modification to h^* to reduce its error bias. To the best of our knowledge, not much work has done on this issue and it is worth of more investigating. Here we present a simple modification algorithm *RandomFlip* that can be applied to any classifier. After the classifier makes predictions, the algorithm randomly flips the predicted labels of some individuals with a certain probability p . We compute p according to the predictions of h^* over \mathcal{D}^* . Denoting $\delta = \tau - |\text{DE}_{\mathcal{D}^*}|$, to achieve $|\text{DE}_{\mathcal{D}^*} + \varepsilon_{h^*, \mathcal{D}^*}| \leq \tau$ it suffices to make $|\varepsilon_{h^*, \mathcal{D}^*}| \leq \delta$. Recall that $\varepsilon_{h^*, \mathcal{D}^*} = \varepsilon_1^+ - \varepsilon_2^+ - (\varepsilon_1^- - \varepsilon_2^-)$. Thus, it suffices to make $|\varepsilon_1^+ - \varepsilon_2^+| \leq \delta/2$ and $|\varepsilon_1^- - \varepsilon_2^-| \leq \delta/2$. Suppose that $\varepsilon_1^+ - \varepsilon_2^+ > \delta/2$. For each individual that is predicted positive, the algorithm randomly changes its predicted label from l^+ to l^- with probability p so that after the change we have $0 \leq \varepsilon_1^+ - \varepsilon_2^+ \leq \delta/2$. The false positive rate ε_1^+ after the change is given by $\frac{fp - p \cdot fp}{tp + fp}$, and the false negative rate ε_2^+ after the change is given by $\frac{fn + p \cdot tp}{tp + fp}$. Refer to Table 2 for the meanings of the notations. Please note that the confusion matrix is built based on the predictions of h^* over the data in \mathcal{D}^* with $C = c^+$. Thus, probability p should satisfy

$$0 \leq \frac{fp - p \cdot fp}{tp + fp} - \frac{fn + p \cdot tp}{tp + fp} \leq \frac{\delta}{2},$$

which results in $\varepsilon_1^+ - \varepsilon_2^+ - \delta/2 \leq p \leq \varepsilon_1^+ - \varepsilon_2^+$. Similar operations apply for other situations.

6 Empirical Example

In this example, we learn a causal model \mathcal{M} from the Adult dataset [Lichman, 2013] using the software Tetrad [Glymour and others, 2004]. We use \mathcal{M} to generate a training dataset \mathcal{D} consisting of 10000 tuples. We treat *sex* as C and *income* as L . Two preprocessing methods are

used for modifying the training dataset: the *Massaging* [Kamiran and Calders, 2009a] that modifies L and the *Disparate Impact Removal* [Adler et al., 2016] that modifies \mathbf{R} . The discrimination is measured to be 0.018 on \mathcal{M} and 0.018 on \mathcal{D} . We assume a rigorous threshold $\tau = 0.01$, i.e., we want to ensure that the discrimination in the prediction is not larger than 0.01.

To show that the discrimination-free predictions cannot be achieved by only modifying the training data, we apply the *Massaging* method to completely remove the discrimination contained in \mathcal{D} . The discrimination on the modified data \mathcal{D}^* is measured to be 0. Then, we build an SVM classifier h^* on \mathcal{D}^* . Finally, we measure the discrimination in \mathcal{M}_{h^*} according to Proposition 3, which equals 0.071. As a result, discrimination still exists in the prediction. Similar results are observed for the *Disparate Impact Removal* method.

To show the effectiveness of the two-phase framework, we first apply the two different methods, the *Massaging* and the *Disparate Impact Removal*, to completely remove the discrimination in \mathcal{D} , obtaining the modified datasets \mathcal{D}_1^* and \mathcal{D}_2^* . Two SVM classifiers h_1^* and h_2^* are built on \mathcal{D}_1^* and \mathcal{D}_2^* respectively. Then, we apply the *RandomFlip* algorithm to both classifiers. Finally, we measure the discrimination in $\mathcal{M}_{h_1^*}$ and $\mathcal{M}_{h_2^*}$. The results shows that $|\text{DE}_{\mathcal{M}_{h_1^*}}| = 0.005$ and $|\text{DE}_{\mathcal{M}_{h_2^*}}| = 0.020$, which validates the correctness of our theoretical results.

7 Related Work

A large number of methods have been proposed for constructing discrimination-free classifiers, which can be broadly categorized as the inprocessing methods that achieve non-discrimination by modifying the classifiers, and the preprocessing methods that achieve non-discrimination by modifying the training data. The methods in the first category usually perform some tweak or develop some regularizer for the classifier to correct or penalize discriminatory outcomes [Kamiran et al., 2010; Calders and Verwer, 2010; Kamishima et al., 2011; Kamishima et al., 2012; Fish et al., 2016]. The regularization terms could not be linked with the non-discrimination threshold τ , and hence cannot achieve non-discrimination guarantee.

Our work falls into the second category that modifies the training data. Previous work has assumed that the predictions would certainly contain no discrimination as long as the training data is modified to be discrimination-free. So they have focused on minimizing the modification to the training data so that the classification accuracy is not decreased too much. Based on whether the proposed methods can ensure non-discrimination in prediction using our proposed framework, they can be further categorized as the methods that only modify the label, including the *Massaging* [Kamiran and Calders, 2009a; Žliobaitė et al., 2011] and the *CBN-Based Removal* [Zhang et al., 2016], and the methods that modify the data other than the label, including the *Preferential Sampling* [Kamiran and Calders, 2012; Žliobaitė et al., 2011], the *Reweighting* [Calders et al., 2009], and the *Disparate Impact Removal* [Feldman et al., 2015;

Adler *et al.*, 2016]. Only the first category can guarantee non-discrimination in prediction based on our framework.

8 Conclusions and Future Work

In this paper, we examined the fundamental assumption made in the preprocessing methods using the causal model. Our theoretical results show that: 1) only removing discrimination from the training data cannot ensure non-discrimination in the prediction for any classifier; and 2) when removing discrimination from the training data, one should only modify the labels in order to obtain a non-discrimination guarantee. Based on the results, we developed a two-phase framework for constructing a discrimination-free classifier with a theoretical guarantee. In the future work, we will apply our results to the study of the trade-off between classification accuracy and the non-discrimination guarantee.

References

- [Adler *et al.*, 2016] Philip Adler, Casey Falk, Sorelle A Friedler, Gabriel Rybeck, Carlos Scheidegger, Brandon Smith, and Suresh Venkatasubramanian. Auditing black-box models for indirect influence. In *Proceedings of ICDM 2016*, 2016.
- [Calders and Verwer, 2010] Toon Calders and Sicco Verwer. Three naive bayes approaches for discrimination-free classification. *Data Mining and Knowledge Discovery*, 21(2):277–292, 2010.
- [Calders *et al.*, 2009] Toon Calders, Faisal Kamiran, and Mykola Pechenizkiy. Building classifiers with independence constraints. In *Data mining workshops, 2009. ICDMW’09. IEEE international conference on*, pages 13–18. IEEE, 2009.
- [Feldman *et al.*, 2015] Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and removing disparate impact. In *KDD*, pages 259–268. ACM, 2015.
- [Fish *et al.*, 2016] Benjamin Fish, Jeremy Kun, and Ádám D Lelkes. A confidence-based approach for balancing fairness and accuracy. In *Proceedings of the 2016 SIAM International Conference on Data Mining*, pages 144–152. SIAM, 2016.
- [Glymour and others, 2004] Clark Glymour et al. The TETRAD project. <http://www.phil.cmu.edu/tetrad>, 2004.
- [Kamiran and Calders, 2009a] Faisal Kamiran and Toon Calders. Classifying without discriminating. In *Computer, Control and Communication, 2009. IC4 2009. 2nd International Conference on*, pages 1–6. IEEE, 2009.
- [Kamiran and Calders, 2009b] Faisal Kamiran and Toon Calders. Discrimination-aware classification. In *21st Benelux Conference on Artificial Intelligence (BNAIC)*, pages 333–334, 2009.
- [Kamiran and Calders, 2012] Faisal Kamiran and Toon Calders. Data preprocessing techniques for classification without discrimination. *KAIS*, 33(1):1–33, 2012.
- [Kamiran *et al.*, 2010] Faisal Kamiran, Toon Calders, and Mykola Pechenizkiy. Discrimination aware decision tree learning. In *ICDM*, pages 869–874. IEEE, 2010.
- [Kamishima *et al.*, 2011] Toshihiro Kamishima, Shotaro Akaho, and Jun Sakuma. Fairness-aware learning through regularization approach. In *ICDMW*, pages 643–650. IEEE, 2011.
- [Kamishima *et al.*, 2012] Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. Fairness-aware classifier with prejudice remover regularizer. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 35–50. Springer, 2012.
- [Koller and Friedman, 2009] Daphne Koller and Nir Friedman. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.
- [Lichman, 2013] M. Lichman. UCI machine learning repository. <http://archive.ics.uci.edu/ml>, 2013.
- [Murphy, 2012] Kevin P Murphy. *Machine learning: a probabilistic perspective*. MIT press, 2012.
- [Pearl, 2009] Judea Pearl. *Causality: models, reasoning and inference*. Cambridge university press, 2009.
- [Romei and Ruggieri, 2014] Andrea Romei and Salvatore Ruggieri. A multidisciplinary survey on discrimination analysis. *The Knowledge Engineering Review*, 29(05):582–638, 2014.
- [Žliobaitė *et al.*, 2011] Indre Žliobaitė, Faisal Kamiran, and Toon Calders. Handling conditional discrimination. In *ICDM*, pages 992–1001. IEEE, 2011.
- [Zhang *et al.*, 2016] Lu Zhang, Yongkai Wu, and Xintao Wu. Achieving non-discrimination in data release. *arXiv preprint arXiv:1611.07438*, 2016.